Title

Overconfidence in Probability Distributions: People Know They Don’t Know but They Don’t Know What to Do About It

Authors

Soll, Jack B
Palley, Asa
Klayman, Joshua
et al.

Publication Date

2023-03-22

Peer reviewed

Overconfidence in Probability Distributions:

People know they don't know but they don't know what to do about it

Jack B. Soll

Fuqua School of Business, Duke University, Durham, NC 27708. jsoll@duke.edu

Asa B. Palley

Kelley School of Business, Indiana University, Bloomington, IN 47405. apalley@indiana.edu

Joshua Klayman

The University of Chicago Booth School of Business, Chicago, IL 60637. fjklayma@chicagobooth.edu

Don A. Moore

Haas School of Business, University of California, Berkeley, CA 94720. dmoore@haas.berkeley.edu

Abstract: Quantifying uncertainty in the form of a probability distribution is a critical step in many managerial decision problems. However, a large body of previous work has documented pervasive overconfidence in subjective probability distributions (SPDs). We develop new methods to analyze judgments about variables which entail both epistemic and aleatory uncertainty and, in three experiments, study the quality of people's SPDs in such settings. We find that although SPDs roughly match the aleatory concentration of the real-world distributions, people's judgments are consistently overconfident because they fail to spread out probability mass to account for their own epistemic uncertainty about the location and other properties of the distribution. Although people are aware of this lack of knowledge, they do not know how to appropriately incorporate it into their SPDs. Our results offer new insights into the causes of overconfidence in real-world judgment domains and shed light on potential ways to address this fundamental bias.

1. Introduction

People are too confident in the accuracy of their beliefs. For example, 90% confidence intervals contain the truth as little as 50% of the time—implying that judges are surer of their knowledge than they deserve to be (for an early review, see Lichtenstein et al. 1982). We refer to this as *overprecision*, which is one of several different types of overconfidence (Moore and Healy 2008). Overprecision arises when judges concentrate too much probability around their favored answer relative to their accuracy, underestimating the probability that the truth may be much farther away. Over the past quarter century, many studies across disciplines such as psychology, decision theory, and finance have shown that subjective probability distributions are typically too narrow and exhibit overprecision (e.g., Juslin et al. 1999, Soll and Klayman 2004, Teigen and Jørgensen 2005, Budescu and Du 2007, Glaser and Weber 2007, Jain et al. 2013). At the same time, Moore, Carter, and Yang (2015) find that, when there is a distribution of values in a population (e.g., possible outcomes of 500 plays of a gamble), subjective probability distributions are less concentrated than the true distribution of outcomes. In this paper, we present new methods that allow us to reconcile this apparent discrepancy. The results offer insight into the psychological processes that lead to overprecision.

The most popular paradigm for studying overprecision asks people to estimate factual quantities about which they are unsure, such as the weight of a Boeing 787 or the year in which Mozart was born. These types of questions entail *epistemic uncertainty*—doubt in the judge's mind about information that is, at least in principle, knowable. The uncertainty arises from the judge having only partial information. For example, the judge may know that Mozart was a classical composer, and that classical composers lived in the eighteenth or nineteenth centuries. In contrast, with chance devices such as random number generators, dice, and coin flips, the best anyone can do is to specify a probability distribution of potential outcomes. For instance, the probability of rolling 7 with a pair of standard dice is 6/36, and the probability of rolling 6, 7, or 8 is 16/36. This is *aleatory uncertainty*—a representation of an outcome as inherently unpredictable, but with a knowable distribution of probabilities across instances.

The two types of uncertainty have been shown to correspond to different reasoning processes, by which people conceptualize an instance either as drawn from a class of events (aleatory) or as a unique and knowable event (epistemic) (Fox and Ülkümen 2011). For example, most people think of a coin flip as an exemplar belonging to a class of possible flips, in which there are two types of outcomes, equally likely. Uncertainty is aleatory because, from the perceiver's point of view, either outcome can potentially occur. In principle, though, the outcome of a coin flip could be represented as a unique event, in which uncertainty is epistemic: It arises from a lack of knowledge of the precise physical forces

operating in the moment on that particular coin. In contrast, most people would consider Mozart's birth date to be a unique instance. The uncertainty here is epistemic because it corresponds to an assessment of one's degree of knowledge. People would likely ask themselves "How much do I know about Mozart?" as opposed to thinking about Mozart's birth date as a random draw from the distribution of birth dates of classical composers. Different types of events may evoke thoughts of aleatory uncertainty, epistemic uncertainty, or both, depending on factors such as repeatability (e.g., Mozart can only be born once) and the recognition of the role of chance in producing outcomes (Nisbett et al. 1983).

In a series of papers, Fox, Ülkümen, and their colleagues examine the distinction between aleatory and epistemic uncertainty (Fox and Ülkümen 2011; Tannenbaum et al., 2017; Ülkümen et al. 2016). It is similar to the distinction between case-based reasoning, which involves thinking about the causal propensities and attributes of a specific target, and class-based reasoning in which people calculate relative frequencies or imagine a distribution of possibilities (Howell and Burnett 1978; Peterson and Pitz 1988; Kahneman and Tversky 1982; Gigerenzer 1994; Teigen 1994). There are, however, two additional features that make the aleatory-epistemic distinction especially well-suited to our purposes. First, epistemic versus aleatory uncertainty captures the distinction between an impression of one's own lack of knowledge and an impression of randomness in the world. Second, a given judgment problem can include elements of each. It is not a dichotomy, but rather a continuum ranging from pure aleatory uncertainty (e.g., a game of chance) to pure epistemic uncertainty (e.g., a trivia question).

Many important decisions involve both epistemic and aleatory uncertainty. How much to save for retirement depends on how long one will live. The decision to invest in a startup venture depends on the probability that the business will survive and its long term-profitability. Whether to undergo a complicated surgery depends on the likelihoods of different potential outcomes for the patient. An individual who needs to specify one of these probability distributions would do well to think about probabilities or relative frequencies within a reference class of similar events, which is an aleatory representation. For instance, when planning for retirement it is useful to know how long others like you have lived. When valuing a startup, it is helpful to know the three-year failure rate of similar entrepreneurial ventures. Yet there are important elements of epistemic uncertainty in these judgments as well, and individuals may have varying degrees of idiosyncratic knowledge about a particular variable. For example, a doctor judging the probabilities of different surgical results may combine specific information from a patient's medical history with knowledge of the base rates of different outcomes. Epistemic uncertainty can also arise because the judge is uncertain about what the relevant probability

distribution is. The medical literature may offer limited or conflicting data about complication rates, leading to uncertainty in the doctor's mind about the base rates of those outcomes.

Studying beliefs about aleatory uncertainty necessitates having a distribution of possible answers against which to compare participants' responses. Rather than focusing on purely aleatory events, we introduce an epistemic component by examining beliefs about everyday domains with which participants might be imperfectly familiar, such as commute times, housing values, and temperatures. These events all have an aleatory component—for example, it is natural to think of a distribution of home prices in Chicago. The epistemic component arises because an individual may not know the distribution for sure, but rather have an imperfect impression of it. This combination of aleatory and epistemic uncertainty contrasts with the kinds of uncertainty studied in most of the prior literature, where epistemic uncertainty predominates and there is no empirical distribution to compare to (to wit, unless we invoke parallel universes, there is no distribution of Mozart's birth date). Furthermore, we introduce a novel methodology that takes advantage of this partition of uncertainty into aleatory and epistemic components, and affords an opportunity to gain new insights into the psychology that underlies overprecision in judgment.

2. Two Standards for Subjective Probability Distributions

In our studies, we ask people to estimate a distribution of probabilities for a member of a class of events, objects, or people. For example, we ask, "If we were to randomly choose one person from Philadelphia with a full-time job, what would be their average daily commute time?" A judge could believe that it is more likely to find a person who commutes between 20 and 30 minutes each way than one who commutes 0 to 10 minutes or 90 to 100 minutes, and so on. We are interested in the question of whether subjective probability distributions (SPDs) like this are, on average, the right shape. That is, are probabilities systematically overly concentrated in a few favored ranges (i.e., too narrow), too dispersed across many ranges (i.e., too wide), or about right? In making such comparisons, two different standards apply.

When the researcher knows the true distribution of probabilities (as we do thanks to data from the U.S. Census Bureau) it is possible to compare the concentration of the reported subjective distribution to that of the empirical distribution. There is limited research on this topic, but two sets of studies suggest that subjective distributions could be more dispersed than the corresponding empirical distributions. Nisbett and Kunda (1985) asked one group of college students to report their own attitudes (e.g., opinion of Ronald Reagan as president) and behaviors (e.g., frequency of going to concerts) and another group to estimate the distribution of one hundred of their peers' answers to

those same questions. They found that the standard deviations of the estimated distributions were on average about 10% greater than the empirical ones. Moore, Carter, and Yang (2015) compared subjective and objective distributions for the outcomes of various randomizing devices. For example, they used a Galton Board wherein a ball dropped from a slot at the top of the machine bounces over a series of staggered pegs to land in a bin at the bottom, producing a binomial distribution. Similar to the Nisbett and Kunda result, subjective distributions for the Galton board as well as other binomial distributions were more dispersed than the objective ones.

Alternatively, we can compare the expressed probabilities of finding a member in a given range or set of ranges to the actual probability of finding a member in that range. That is, the judge might assign 25% probability to finding a commuter in the 20-30 range, when in fact only 15% of commuters fall in that range. As noted earlier, a large body of previous research is consistent in finding that subjective judgments tend to be overprecise. That is, judges tend to express too much confidence that the target estimate will fall within a given range. Using this standard with their randomizing devices, Moore, Carter, and Yang (2015) found that participants' confidence intervals were too narrow. In other words, they observed that SPDs fell in between two standards—they were wider than the objective distribution, but not wide enough to be well calibrated.

Moore et al. interpreted these results as paradoxical, but the difference can be understood in terms of aleatory and epistemic uncertainty. In judging subjective probabilities of a range of possible events, people should consider both. They should consider that, within a class of events, individuals vary along any given measure (e.g., people have different incomes), and they should also consider that they have less-than-perfect knowledge of what that distribution is. People may misestimate the extent to which exemplars are concentrated in a narrow range vs. spread across a wide range (e.g., the degree of income inequality in a given city). They may also misestimate the central tendency of the distribution (e.g., the median income in that city). This combination of aleatory and epistemic uncertainty means that, to be well-calibrated, a judge must provide a wider distribution than the one the U.S. Census Bureau knows (with minimal epistemic uncertainty). If judges respond to epistemic uncertainty, but insufficiently, they might very well provide SPDs that are wider than the empirical distribution, but too narrow to be well calibrated (for a related explanation, see Camilleri and Newell 2019).

3. The Present Research

We introduce a new measure, *concentration,* by which to compare subjective probabilities to each of these two standards. Concentration, which we define formally in the next section, measures the extent to which probability mass piles up in one part of the spectrum of possible outcomes versus being

spread across possible answers. We say that a subjective distribution is sub-concentrated if it is less concentrated than the objective or empirical distribution. Super-concentration reflects greater concentration in the subjective than objective distribution. Roughly speaking, sub-concentration means that a judge's SPD is more spread out across possible answers compared to the actual distribution of outcomes.

We examine several explanations, not mutually exclusive, for how overprecision might emerge from how people perceive and combine aleatory and epistemic uncertainty. First, people may believe that distributions of outcomes in the world are more concentrated than they really are—their assessments of aleatory uncertainty are too narrow. Second, people may believe that they know the empirical distribution better than they really do—they perceive less epistemic uncertainty than they should in order to be well-calibrated. Consistent with this, many authors have argued or implied that overprecision is the direct result of respondents' failure to appreciate or to admit how much they do not know. When Alpert and Raiffa (1982) documented the low hit rates of subjective confidence intervals, they implored their subjects, "For heaven's sake, *Spread Those Extreme Fractiles*! Be honest with yourselves! Admit what you don't know!" (p. 301, emphasis in original). This quote also captures a third explanation. Perhaps people do perceive and appreciate epistemic uncertainty, but don't recognize that they should therefore "Spread Those Extreme Fractiles!", or they do so insufficiently.

Testing for these three explanations requires us to develop better methods to assess the concentration of SPDs. To do this, we combine methods from two distinct domains, overconfidence research and population economics. Research in subjective confidence often has asked for X% subjective confidence intervals. However, these provide limited information about the full distribution. An alternative is the SPIES (Subjective Probability Interval EstimateS) method (Haran, Moore, and Morewedge, 2010), in which people estimate probabilities for given intervals, rather than reporting the interval size corresponding to a fixed probability (see also Goldstein & Rothschild, 2014). The SPIES method permits researchers to elicit detailed probability distributions by asking participants to assign probabilities to a set of mutually exclusive and exhaustive outcome bins. Haran et al. (2010) show this method to produce better-calibrated subjective distributions, compared to earlier methods eliciting a single interval corresponding to a fixed probability. After eliciting distributions using the SPIES method, we repurpose a pair of well-established statistics used by population economists to measure the concentration of a resource across a population, namely the Lorenz curve and its numerical summary, the Gini coefficient (Gastwirth, 1972; Gini, 1912; Lorenz, 1905). We compare the concentration of SPDs with both the concentration of the real-world target distribution and the concentration needed to be

well-calibrated. Having a single measure by which to make both comparisons will allow us to distinguish among different potential explanations.

4. Measures of concentration and calibration

To analyze the results of our experiments we need appropriate measures of the concentration and calibration of an individual's subjective distributions. The standard deviation of the subjective distribution might seem a straightforward measure of concentration. However, when judgments are grouped into categories, estimates of standard deviation depend heavily on assumptions about the distribution of values within categories, especially unbounded end-categories ("greater than __", "less than __"), and on the shape of the distribution. The Lorenz curve and the Gini coefficient are well suited to measuring concentration when such details are not knowable. Whereas economists use the Lorenz curve to plot the degree to which a resource such as wealth is concentrated in a few hands or many, we use it to plot the degree to which a judge piles up probability mass in a few categories or many.

In our studies, judges assess probabilities for each of a set of mutually exclusive and exhaustive ranges, or *bins*. To plot Lorenz curves, we subdivide the horizontal axis from 0 to 1 into equal fractional increments, with each increment adding another bin cumulatively. By definition, Lorenz curves always begin at (0, 0) and end at (1, 1). Between those endpoints, the first increment in our plots represents the bin assigned the most probability, the first two increments represent the two bins assigned the most probability, and so on. The notation 3/5, for example, indicates the cumulative result for top three bins out of five. The vertical indicates the cumulative probabilities assigned to each subset of bins. We compare three different Lorenz curves and associated Gini measures, the details of which are best explained through an example.

Figure 1 provides an example of Lorenz and Gini calculations for an individual who has provided probability judgments for the likely finishing time of an individual chosen at random from among all those who completed the 2016 Boston Marathon, which is the same as the distribution of finishing times. The individual's responses appear in orange in the upper left panel. This judge estimated finishing times between 5:00 and 5:59 hours to be most likely, assigning a subjective probability of 0.38 to this event, and finishing times under 3:00 hours to be least likely, assigning this event a subjective probability of 0.05.

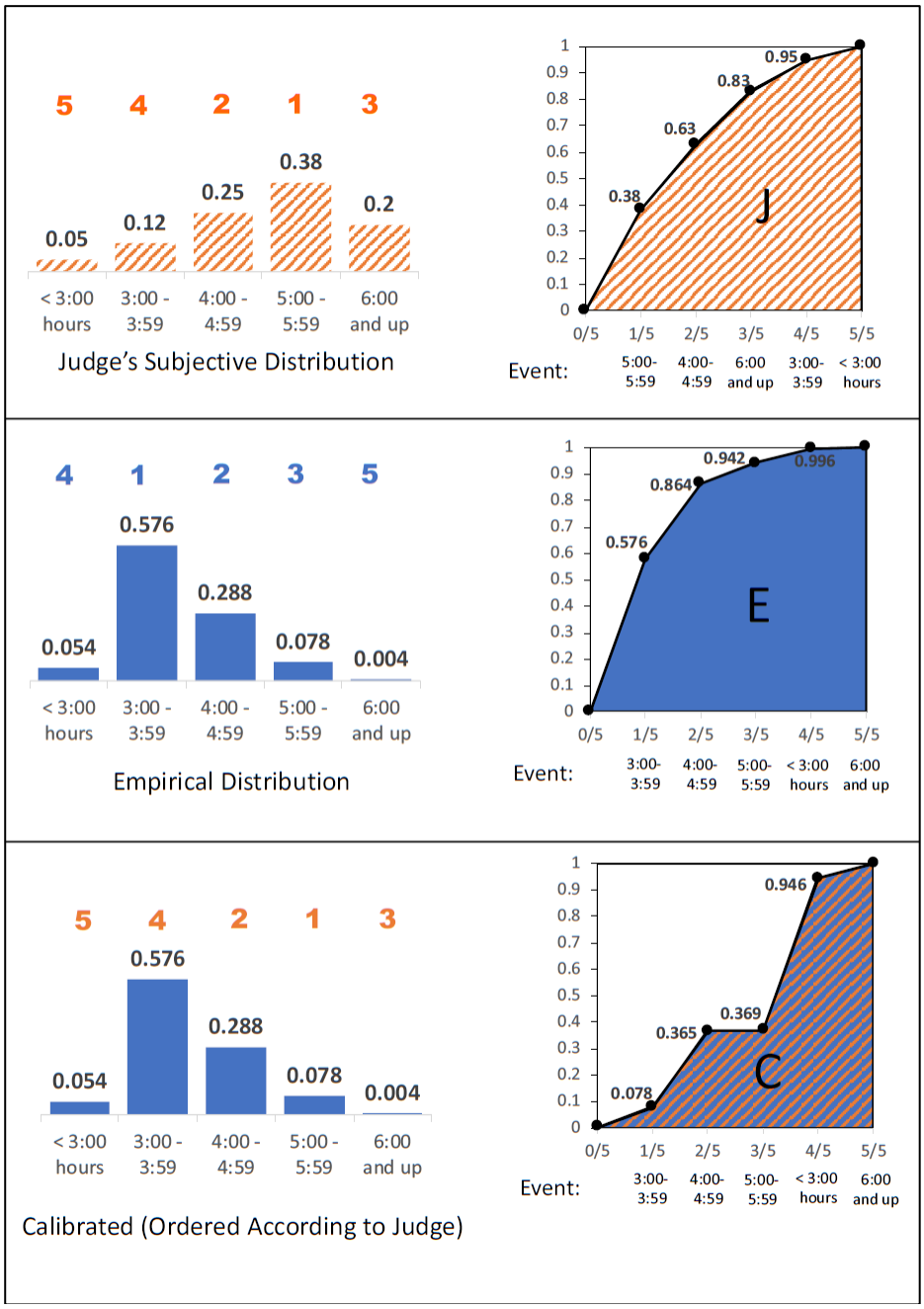

Figure 1. Subjective, empirical, and calibrated Lorenz curves for a hypothetical judge estimating the distribution of finishing times in the 2016 Boston Marathon.

The judge's Lorenz curve, plotted in the top right panel, is constructed by successively cumulating the subjective probabilities of the ranges, starting with the one judged most likely (5:00 to 5:59) and ending with the one judged least likely (< 3:00). The steepness of the judge's curve represents the degree to which the judge believes that some ranges are more likely than others. In the least concentrated possible distribution, each bin is assigned equal probability and the curve follows the identity line from (0,0) to (1,1). The judge's Gini coefficient is given by $G^*_{Judge} = \frac{N}{N-1}(2J - 1)$, where *J* is the area under the judge's Lorenz curve and *N* is the number of events.[1] G^*_{Judge} equals 0 when each bin is assigned equal probability and equals 1 when all probability is assigned to a single bin. In the Boston Marathon example, *N = 5* and *J = 0.658*, yielding $G^*_{Judge} = 0.395$. offers several advantages for our present purposes. First, it provides a pure measure of a distribution's concentration independent of hit rates and accuracy. Second, by summarizing concentration into a single number in units of probability, Gini coefficients facilitate comparisons that speak directly to overprecision. Third, it is easy to compute with SPIES elicitations.

Given data on the frequencies of each event, we can construct a Lorenz curve representing the observed distribution in the real-world data. This empirical Lorenz curve is generated by ordering the events from most to least likely according to their observed frequencies while cumulating these empirical probabilities. Letting *E* denote the area under this curve, the empirical Gini coefficient is given by $G^*_{Empirical} = \frac{N}{N-1}(2E - 1)$. The middle panel of Figure 1 displays the observed frequencies of finishing times in the 2016 Boston Marathon. Finishing times between 3:00 and 3:59 hours were most frequent, with 57.6% of runners falling into this bin. Finishing times over 6:00 hours were least frequent, with only 0.4% of runners falling into this bin. The corresponding empirical Lorenz curve is constructed by successively cumulating the observed event frequencies in decreasing order, yielding *E = 0.776* and $G^*_{Empirical} = 0.689$. The quantity $G^*_{Judge} - G^*_{Empirical}$ provides a measure of the extent to which the judge's SPD is more or less concentrated than the observed distribution of events. For this judge, this difference is -0.294, with the negative sign indicating that the judge's SPD is sub-concentrated—less concentrated than the observed distribution of finishing times.

Finally, we consider a third Lorenz curve, constructed by cumulating the empirical frequency of events that fall into each bin *as they were ordered by the judge*. This calibrated Lorenz curve, displayed

[1] Note that we apply an adjustment, $G^* = \frac{N}{N-1}G$, where G is the standard Gini. This is needed in domains in which the population is smaller, such as concentration of market share among companies (e.g., Collins & Preston, 1961) and concentration of crime in particular neighborhoods (Bernasco & Steenbeek, 2017). This correction keeps the Gini always bounded by 0 and 1 (Deltas, 2003).

in the lower panel of Figure 1, depicts the confidence levels that would be assigned to each of the judge's cumulative bins by an imaginary judge named *Calibra*[2] who knows the empirical distribution perfectly. The associated Gini coefficient is $G^*_{Calibra} = \frac{N}{N-1}(2C - 1)$, where *C* is the area under the calibrated Lorenz curve. Unlike the previous two measures, $G^*_{Calibra}$ can take negative values, ranging from -1 to 1. The probabilities in our example yield an area of *C = 0.452* and $G^*_{Calibra} = -0.121$. The quantity $G^*_{Judge} - G^*_{Calibra}$ provides a measure of the difference between the confidence the judge expressed in their subjective judgments and the level of confidence they were entitled to hold given the event order they had expressed. A positive difference indicates *overprecision*, meaning that the implied confidence intervals generated by sets of events that the judge deemed most likely were too narrow. For this judge, $G^*_{Judge} - G^*_{Calibra} = 0.516$.

Comparing subjective, empirical, and calibrated Gini coefficients corresponds to, and extends, other well-established measures of overconfidence. For example, we can use these differences to calculate a measure of *global overprecision,* representing the difference between the judge's subjective probabilities and the empirical probabilities, averaged over the judge's top 1, top 2, …, and top *N* - 1 cumulated categories. In the Boston Marathon example, overprecision for the judge's top category is 0.380 – 0.078 = 0.302, for the top two categories is 0.630 – 0.365 = 0.265, and so on. If we average the four results for overprecision calculated in this manner, we find that global overprecision equals 0.258. More directly, global overprecision can be equivalently calculated according to $\left(G^*_{Judge} - G^*_{Calibra}\right)/2$.

These measures can be used across a variety of variable types (e.g., binary, continuous, multiple choice, ordinal) and elicitation formats. One caveat most be noted, however: The Lorenz curves can vary depending on the number of ranges used to divide up a continuous scale, and on where the "greater than ___" and "less than ___" end-categories begin. In the studies that follow, we mitigate this concern by comparing coefficients from comparable partitions and by using pre-existing, externally determined partitions when possible.

5. Experiment 1

Our first experiment had two main goals: The first goal was to test the generality of the finding by Moore, Carter, and Yang (2015) that judges' subjective probabilities are less concentrated than the empirical distribution, but more concentrated than necessary for good calibration. The second goal was to test whether this pattern can be understood as a directionally correct, but insufficient, response to

[2] The legend of Calibra was told by Soll and Klayman (2004).

epistemic uncertainty about the empirical distribution. Pre-registration materials for all of our experiments are available at https://osf.io/dt7cq. For Experiment 1 we pre-registered four hypotheses:

- H1: $G^*_{Judge} < G^*_{Empirical}$. As observed in the Nisbett and Kunda (1985) and Moore et al. (2015) studies, SPDs are less concentrated than the empirical probability distribution.
- H2: G^*_{Judge} is higher when epistemic uncertainty is lower. SPDs are more concentrated when judges have more information about the objective distribution (e.g., its median or mode).
- H3: $G^*_{Judge} > G^*_{Calibra}$. Based on the well-established finding of pervasive overprecision, SPDs are more concentrated than they would need to be for good calibration.

We hypothesized this pattern because we expected judges to respond to epistemic uncertainty in the directionally-correct way, by widening their SPDs in comparison to the empirical (H1 and H2), but they do that insufficiently, leaving them with SPDs that are still not wide enough to be well calibrated (H3). For many versions of "insufficiently," we would expect the gap between $G^*_{Calibra}$ and G^*_{Judge} to be bigger when the gap between $G^*_{Calibra}$ and $G^*_{Empirical}$ is bigger. Thus:

- H4: Overprecision, measured by $G^*_{Judge} - G^*_{Calibra}$, is larger the greater the mismatch between the empirical distribution and the well-calibrated one, as measured by $G^*_{Empirical} - G^*_{Calibra}$.

To understand H4, note that for a randomly chosen stimulus it must be the case that $G^*_{Calibra} \leq G^*_{Empirical}$. A person with perfect knowledge of the empirical distribution would simply report that same distribution, so a judge with less knowledge would need to be less concentrated than that in order to be well-calibrated. The gap between $G^*_{Calibra}$ and $G^*_{Empirical}$ therefore measures how well one knows the empirical distribution, and H4 posits that those who have less knowledge about the distribution are more overprecise. This is a version of the hard-easy effect in the overconfidence literature (Klayman et al. 1999), which says that lesser knowledge (i.e., harder questions) corresponds to greater overconfidence.

Because we were concerned about participants' understanding of stochastic devices like the Galton Board used by Moore, Carter, and Yang (2015), we drew questions from five diverse domains of everyday knowledge. We approximated a representative sample of questions by selecting items randomly from well-defined populations in each domain. For example, one of the questions asked about the income of a randomly drawn household from a given city, selected at random from the 40 large U.S. cities in the Census Bureau dataset. We did this to reduce concerns that observed overconfidence could be attributable to the over-representation of tricky "contrary" questions—ones for which usually-valid information or intuition points to an incorrect answer (Klayman et al. 1999).

We introduced an experimental manipulation of epistemic uncertainty by telling some participants the medians of the population distributions (e.g., the median household income in Cleveland). We assume that this information reduces epistemic uncertainty, because one source of epistemic uncertainty is not knowing where to locate the distribution (and being aware of that). If people respond appropriately to epistemic uncertainty, their probability distributions should be more concentrated when they know the median than when they do not.

5.1 Methods

5.1.1. Participants. Based on a pre-registered power analysis, we aimed to recruit a sample size of 600 participants from Amazon Mechanical Turk, split evenly between two conditions. Of the 973 people who began the online survey, 324 failed to successfully complete the training, either by dropping out before completing the training and being assigned to a condition (127) or by completing the training but failing to meet the criterion for passing (197). Another 55 passed the training but failed to complete the study. Of the 252 excluded after completing training (197 + 55), 138 were in the provide-median condition and 114 in the no-median condition. This left a final sample of 594. Participants received a base payment of $0.50 and an average bonus of $0.55 for accuracy.

5.1.2. Materials. Figure 2 shows what participants saw. They reported their subjective likelihoods by adjusting the slider bars, which did not need to add up to 100. Rather, participants adjusted the bars to indicate the relative chances of an observation being in that category (e.g., a bar three times as long means that it is three times as likely). We normalized reported distributions by dividing each bin's assigned likelihood by the total across all bins.

For each domain, the spectrum of possible answers was divided into a modest number of response categories, as with the eight categories shown in Figure 3. As we noted earlier, Gini coefficients can be sensitive to the widths and number of categories. Accordingly, the number of categories in each domain was held constant. Four of the domains use the 7 to 12 categories by which the U.S. Census Bureau reported data from the American Community Survey; for the fifth domain, average daily high temperatures, we defined categories in increments of 10 from 0 to 100 Fahrenheit, with end-ranges of "less than or equal to 0" and "greater than 100."

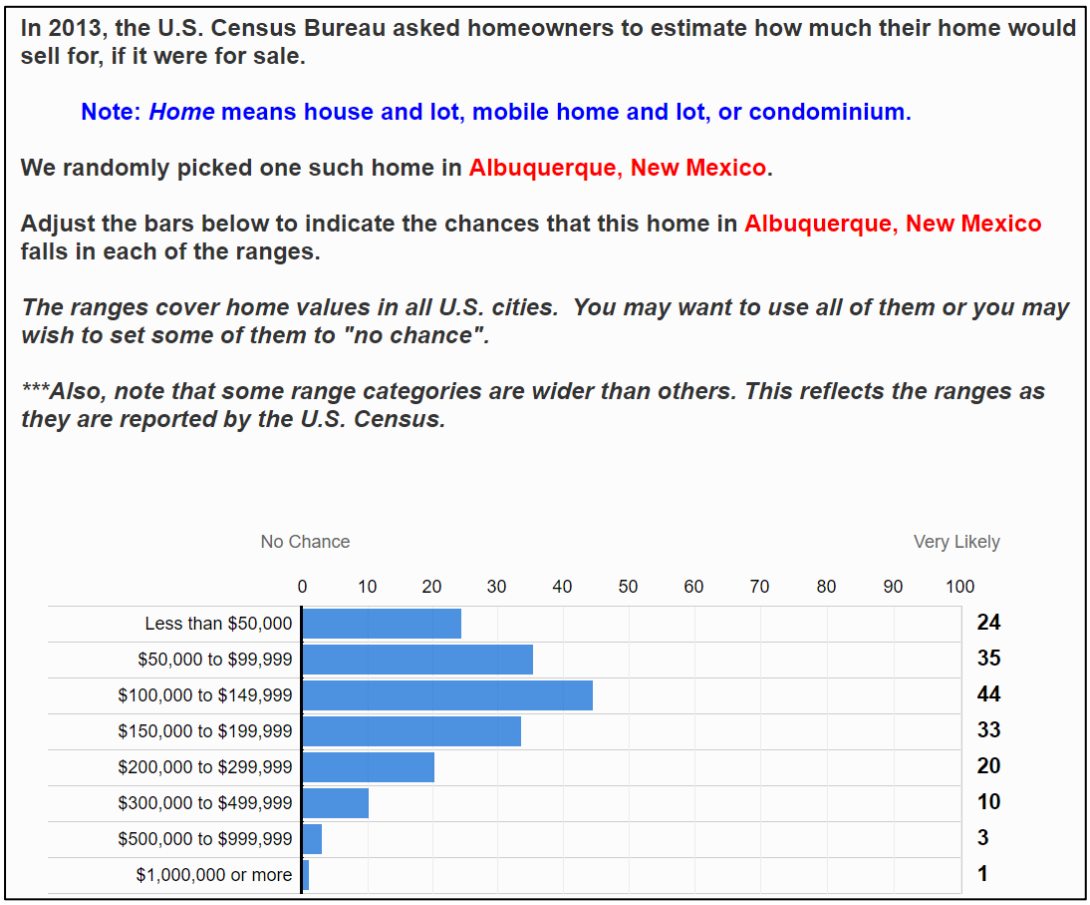

Figure 2. Example question with responses from Experiment 1.

5.1.2. Procedure. A practice item taught participants how to use and interpret the slider bars. They then took a 3-item quiz to make sure that they understood that longer bars represented greater chances, that the relative lengths of bars represented relative chances, and that the bars did not need to sum to 100. Participants had two tries to correctly answer each question, and had to answer all three questions correctly in order to be included in the analysis. In addition, participants who were provided with medians received a brief explanation of the median and a quiz question to test their understanding. However, to ensure that we could generalize the sample in each condition to the same population of participants, we did not exclude any participants based on this question. Exclusions left 310 participants in the experimental condition in which we provided them with medians and 284 in the control condition.

Participants next learned that they could maximize their expected bonus payments by setting the bars to reflect their true beliefs. We told participants to “set the bars to reflect your true beliefs about the relative chances that a random observation will fall in each category. The more accurate your responses, the higher your bonus will be.” We did not provide them with the details of the payoff

formula.[3] After the instructions and comprehension questions, participants saw one question from each of the five domains, presented in random order. For each question, and separately for each participant, a new city was randomly selected from all cities in the database, with the constraint that no city would be duplicated across the five questions. At the conclusion of the study, participants were debriefed and informed that we would deliver base payments within 24 hours and bonuses within one week.

5.2 Results

For each question provided to each participant, we calculated three Gini coefficients: G^*_{Judge}, $G^*_{Empirical}$, and $G^*_{Calibra}$. The means of these coefficients are presented in Table 1 and the results of planned analyses are shown in Table 2. The corresponding Lorenz curves (averaged across judges) appear in Figure 3. We hypothesized that judges' SPDs would be less concentrated than the corresponding empirical distributions (H1), but more concentrated than a well-calibrated judge's would be (H3). As shown in Tables 1 and 2, only the latter prediction is borne out. Global overprecision (the difference between G^*_{Judge} and $G^*_{Calibra}$ divided by 2) averages 0.10. This means that, averaging across all of the nontrivial points on the Lorenz curves (i.e., the top category, the top two categories, …, and the top *N*-1 categories), participants reported confidence that averaged about 10 percentage points higher than their accuracy. Moreover, the judges' average Lorenz curves in Figure 3 lie entirely above the calibrated curves for every domain. This means that participants overestimated the probability at every level of cumulation—they overestimated the chance of finding an instance in what they thought was the most-likely category, the two most likely, and so on, for each domain. Contrary to prediction, though, SPDs were also slightly more concentrated than the empirical distributions, rather than less. However, the average difference between G^*_{Judge} and $G^*_{Empirical}$ is small (.031), and $G^*_{Judge} > G^*_{Empirical}$ in only three of the five domains—commutes, education, and incomes. This is apparent in Figure 3, where for those three domains the judge's Lorenz curve sits entirely above the empirical curve.

[3] Bonus payments were calculated for each question using an incentive-compatible extension of the widely-used Brier (1950) score. For each of the *N* categories *c* available for the question, we calculated the quadratic score the judge would receive if a randomly-chosen instance were to fall in that category: $B_c = \sum_{i=1}^{N} (\boldsymbol{I}_i^c - \hat{p}_i)^2$, where $\hat{p}_i$ is the probability the participant assigned to category *i* and the indicator $\boldsymbol{I}_i^c$ equals 1 when *i* = *c* and 0 otherwise. We then multiplied the quadratic score for each category *c* by the empirical probability p_c that a randomly-chosen member of the population would in fact fall in that category and summed those products to arrive at an average score for that question: $EB = \sum_{c=1}^{N} p_c B_c$. The bonus earned was $20(1 - EB)$ cents for each question.

Table 1. Gini coefficients in Experiment 1.

Domain	G^*_{Judge}	$G^*_{Empirical}$	$G^*_{Calibra}$
Medians not Provided (n = 310)			
Commutes	0.367	0.309	0.088
Education	0.411	0.346	0.161
Home Prices	0.473	0.476	0.243
Incomes	0.418	0.282	0.187
Temperatures	0.514	0.600	0.380
Mean	**0.437**	**0.403**	**0.212**
Medians Provided (n = 284)			
Commutes	0.386	0.311	0.161
Education	0.394	0.344	0.172
Home Prices	0.464	0.488	0.340
Incomes	0.401	0.282	0.198
Temperatures	0.520	0.603	0.461
Mean	**0.433**	**0.406**	**0.266**
Overall Mean	**0.435**	**0.404**	**0.238**

Table 2. ANOVA results for differences in concentration and precision in Experiment 1.

$G^*_{Judge} - G^*_{Calibra}$

Effect	effect df	adjusted effect df	error df	adjusted error df	F	Sig.	Partial Eta Squared
a. Mean	1		592		1794.412	<.001	.752
b. Domain	4	3.130	2368	1852.8	66.491	<.001	.101
c. Information	1		592		41.286	<.001	.065
d. Domain x Information	4	3.130	2368	1852.8	5.077	<.001	.009

$G^*_{Judge} - G^*_{Empirical}$

Effect	effect df	adjusted effect df	error df	adjusted error df	F	Sig.	Partial Eta Squared
e. Mean	1		592		58.580	<.001	.090
f. Domain	4	3.613	2368	2138.9	257.001	<.001	.303
g. Information	1		592		0.649	.421	.001
h. Domain x Information	4	3.613	2368	2138.9	2.464	.049	.004

Note: Adjusted degrees of freedom reflect the Greenhouse-Geisser correction for violations of sphericity; the resulting p-values, reported here, are slightly more conservative.

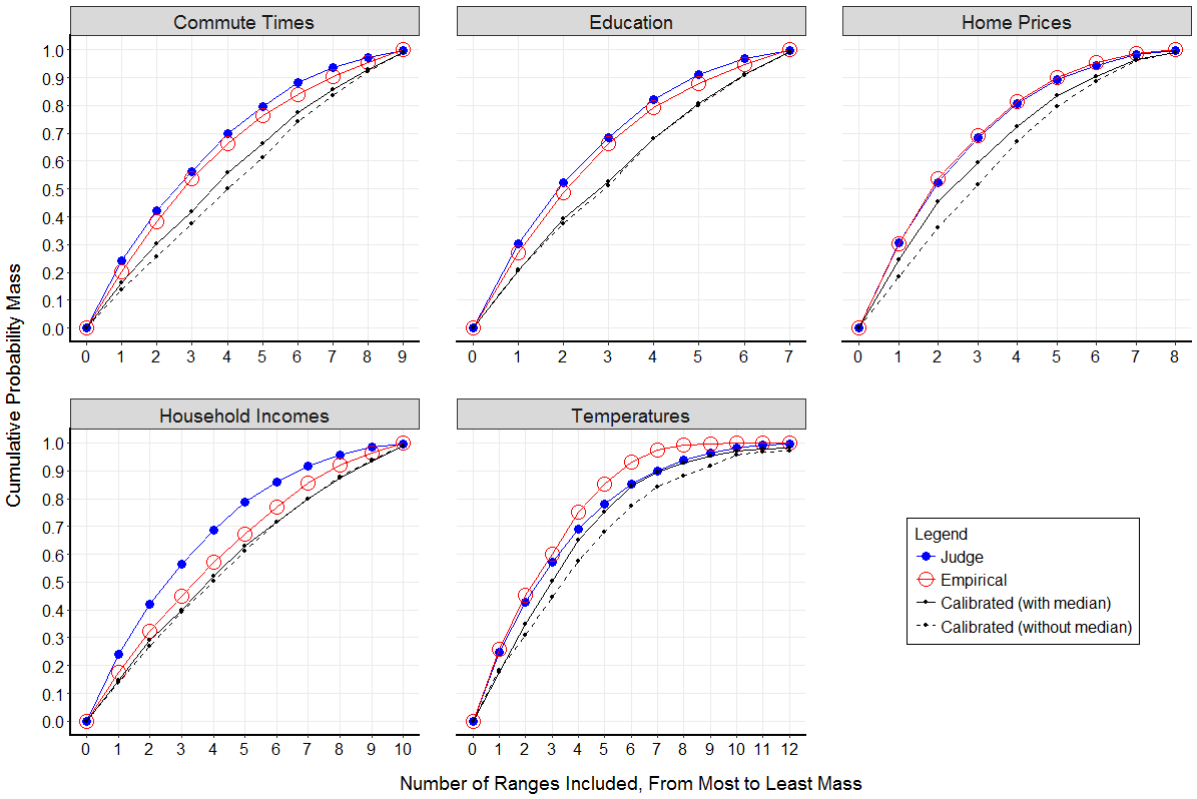

Figure 3. Lorenz Curves from Experiment 1

We also hypothesized that providing participants with the medians would lead to reduced epistemic uncertainty, which would manifest in more concentrated SPDs (higher values of G^*_{Judge}). However, as shown in Table 1, G^*_{Judge} was nearly the same when participants were told the median and when they were not. At the same time, providing the median produced higher $G^*_{Calibra}$. That implies that providing the median did, in fact, reduce epistemic uncertainty: Judges who were given the median were more accurate in centering their reported distributions. Because judges with more information were more accurate but did not give more concentrated SPDs, they were better calibrated than their less-informed counterparts.

We analyzed the Gini coefficients with two planned contrasts: $G^*_{Judge} - G^*_{Empirical}$, which tests how the concentrations of participants' SPDs compare to the concentrations of the underlying empirical distributions, and $G^*_{Judge} - G^*_{Calibra}$, which tests how the SPDs compare to their corresponding well-calibrated distributions. We analyzed each of these contrasts using a 5 (domain) x 2 (information condition) repeated measures ANOVA, with domain being a within-participant variable and information condition between-participants, representing whether the participant learned the medians of the distributions. Results of the two analyses, shown in Table 2, confirm the patterns observed in Table 1:

a) Contrary to Hypothesis 1, judges' distributions are slightly more concentrated than the empirical, although the difference between G^*_{Judge} and $G^*_{Empirical}$ varies strongly across domains.
b) Contrary to Hypothesis 2, providing the median had little effect on judges' relative concentrations, with little variation across domains.
c) In accord with Hypothesis 3, judges' SPDs are substantially more concentrated than they should be to achieve good calibration, although the degree of overprecision varies with domain.
d) In accord with Hypothesis 4, providing the median reduces overprecision, with some domain differences in the degree to which this is true. However, this is not because judges responded insufficiently to epistemic uncertainty. Rather, judges' accuracy increased when given additional information while the concentration of their SPDs remained the same.

5.3 Discussion

In accord with much prior research, Experiment 1 shows that people are consistently overprecise, meaning their SPDs concentrate too much probability in too little of the spectrum. To be well calibrated, SPDs must be wider than the underlying empirical distribution, because they must reflect both the variability in the empirical distribution (aleatory uncertainty) and the likelihood of errors in estimating what that distribution is (epistemic uncertainty). Our results show that, across a variety of domains, SPDs are, on average, slightly narrower than their corresponding empirical distributions. More importantly, there are large differences among domains, suggesting that it is better to think of sub- or super-concentration as a characteristic of a specific domain of judgment rather than as any pervasive tendency.

The experiment also sheds light on why SPDs are not sub-concentrated, as is necessary for good calibration. In the introduction, we posited three explanations: (a) On average, people might believe empirical distributions to be more concentrated than they really are; (b) people might be epistemically too certain–thinking they know more about the empirical distribution than they do; and (c) although epistemic uncertainty would demand that people widen their SPDs, people fail to do so. Our results are most consistent with the third explanation. On average, people are approximately unbiased in their impressions of how dispersed or concentrated outcomes are in the world; however, they do not understand that epistemic uncertainty means their SPDs should be wider than those impressions. Not only do participants give SPDs that are no wider than the empirical distribution, but when we manipulate epistemic uncertainty it has little or no effect on SPDs. That said, it is worth obtaining more direct evidence about whether people respond to epistemic uncertainty and, if so, how. People may

have intuitions about the effects of epistemic uncertainty that Experiment 1 failed to bring to mind, perhaps because the manipulation of epistemic uncertainty was between subjects. We designed Experiments 2 to make differences in epistemic uncertainty more salient and easier to apply.

6. Experiment 2

One reason why participants in our first experiment did not change their distributions in response to epistemic uncertainty may be that they did not think about it when assessing the distributions. In this study, we sought to increase the salience of epistemic uncertainty by varying, within subject, whether we provided the mode of the distribution. All participants estimated probabilities for a randomly selected exemplar drawn from each of six domains, in a procedure similar to that of Experiment 1. One group of participants was told the modal category for each of the six domains (*mode-mode)*; another group was not given that information (*nomode-nomode)*. A third group received the mode for the first three exemplars, but not for the last three (*mode-nomode)*; a fourth group encountered the reverse pattern (*nomode-mode)*. We predicted that losing or gaining information would make participants more aware of their state of knowledge following the change. Thus, if participants respond appropriately to epistemic uncertainty, but need to be prompted to think of it, an obvious change in available information should cue them to reduce the concentration of their subjective distributions when information is removed and to increase the concentration when information is added.

For this study, we modified the materials of Experiment 1 in two ways intended to increase participants' awareness to epistemic uncertainty. Instead of providing the median as in Experiment 1, we provided the mode as additional information. Arguably, the mode is more useful because it tells participants the most likely bin, whereas the median provides only a strong hint about which one it might be. Second, we asked participants for their "confidence" for each bin as opposed to for its "chances." Because people associate the term "confidence" with epistemic uncertainty (Tannenbaum, Fox, and Ülkümen, 2017), specifically asking for this might prime participants to account for it.

6.1 Method

6.1.1. Participants. In this study, participants were assigned to a condition only if they successfully passed the training. Of the 1,166 participants from Amazon Mechanical Turk who began the online survey, 430 failed to successfully complete the training, either by dropping out (79) or by failing to meet the criterion for passing (351). Another 23 passed the training but failed to properly complete the study (7, 6, 6, and 4 in the mode-mode, mode-nomode, nomode-nomode, and nomode-mode

conditions, respectively). In accord with our pre-registered design, analyses included the first 2 qualified participants for each of the 84 unique stimulus sets in each of the 4 conditions, for a total sample of 672.

6.1.2. Design. Experiment 2 had a 2 (Round) x 4 (Information Condition) mixed factorial design. Round was a within-subjects factor. There were two rounds, each with three domains. We systematically varied the order of the six different domains using a Latin Square. Each of the six orderings was duplicated seven times using a different set of randomly selected cities for the six domains, subject to the restriction that each city appear 12 or 13 times and never twice in the same set. We then duplicated those 42 sets, interchanging the rounds (i.e., questions 1-3 became 4-6, and vice versa), creating 84 sets of 6 questions each. Information condition varied between subjects: Participants received information about the mode in Round 1 or did not, and received information about the mode in Round 2 or did not.

6.1.3. Materials and Procedure. We drew questions from the five domains used in Experiment 1, and added a sixth domain, the age of a randomly chosen individual in a selected city. Participants received similar training to those in Experiment 1. They were informed that they could maximize their bonus by setting the bars to reflect their true degree of confidence. Payments were incentive compatible; participants received a base payment of $0.75 and an average bonus of $0.67 for accuracy. Base payments were paid within 24 hours of completing the survey, and bonuses were paid within one week.

At the beginning of Round 1 of questions, participants who received the mode were told, “For the first three items we will provide you with some helpful information. We will tell you the most common category for the given city. Please click below to go to the first item.” Between Rounds 1 and 2 they were told either “For the next three items we will continue to tell you…” (mode-mode condition) or “For the next three items we will no longer provide you with the additional information. We will not tell you…” (mode-nomode condition). For participants who were not given the mode, Round 1 began with just the instruction, “Please click below to go to the first item.” Those in the nomode-nomode condition received no additional instruction at the start of Round 2. For those in the nomode-mode condition, Round 2 was introduced with the same instructions given at the beginning of the mode-mode condition, “For the next three items we will provide you with some helpful information….”.

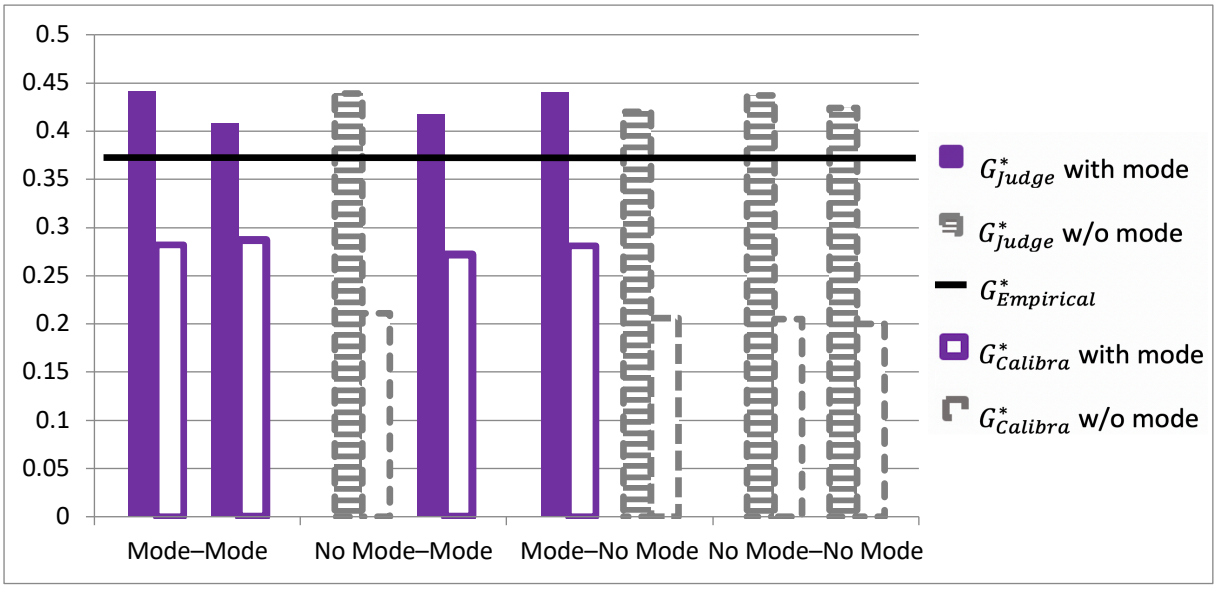

Figure 4. Gini coefficients in Experiment 2. Filled bars display G^*_{Judge}, hollow bars display $G^*_{Calibra}$. The solid horizontal line displays $G^*_{Empirical}$, which is constant across conditions because topics and cities were perfectly balanced. Purple bars with a solid border indicate responses with mode information provided to participants, gray bars with a patterned border indicate responses without mode information.

Table 3. Repeated-measures ANOVA results for differences in concentration in Experiment 2

$G^*_{Judge} - G^*_{Empirical}$

Effect	effect df	error df	F	Sig.	Partial Eta Squared
Mean	1	668	290.58	<.001	.303
Information condition	3	668	0.28	.842	.001
Round	1	668	21.69	<.001	.031
Round x Condition	3	668	0.81	.489	.004

$G^*_{Judge} - G^*_{Calibra}$

Effect	effect df	error df	F	Sig.	Partial Eta Squared
Mean	1	668	2541.08	<.001	.792
Information Condition	3	668	24.92	<.001	.101
Round	1	668	8.91	.003	.013
Round x Condition	3	668	22.34	.001	.091

6.2 Results

As in Experiment 1, we calculated G^*_{Judge}, $G^*_{Empirical}$, and $G^*_{Calibra}$ for each participant's responses in each domain. As before, we also analyzed two difference scores: $G^*_{Judge} - G^*_{Empirical}$, which compares the participant's concentration to the empirical distribution, and $G^*_{Judge} - G^*_{Calibra}$, which compares the judge's concentration to the concentration needed for good calibration. For each measure, we averaged across the three domains in each round. The fact that the design was perfectly balanced meant that the average $G^*_{Empirical}$ was 0.376 in all conditions and rounds. The average Gini coefficients are displayed in Figure 4 and the results of a repeated-measures ANOVA for differences in concentration are presented in Table 3.

As in Experiment 1, participants were on average super-concentrated: $G^*_{Judge} - G^*_{Empirical}$ was significantly positive (*M* = 0.026). As shown by the solid bars in Figure 4, participants were less concentrated in Round 2 than in Round 1, but there were no effects of information condition.

Overall, participants were also overprecise: $G^*_{Judge} - G^*_{Calibra}$ was significantly positive (*M* = 0.185). This effect varied by condition and round. As shown in Figure 4, these findings reflect a consistent effect: There is a small, but statistically significant, main effect of round because, as noted earlier, judges were more concentrated in Round 1 than Round 2, whereas $G^*_{Calibra}$ was virtually the same (.245 and .241, respectively). There is less overprecision when the mode was provided. This is represented by the Round x Condition interaction: The effect of round depended on the presence or absence of information in that round. We followed up on the interaction with separate comparisons of $G^*_{Judge} - G^*_{Calibra}$ for each round. In both rounds, $G^*_{Judge} - G^*_{Calibra}$ was greater without the mode than with it (*M* = 0.230 vs. 0.159, $t(670) = 7.29$, $p < .001$, Cohen's $d = .562$, in Round 1 and *M* = 0.220 vs. 0.132, $t(670) = 9.31$, $p < .001$, Cohen's $d = .718$, in Round 2).

6.3 Discussion

The overall results are consistent with those of Experiment 1. Participants' distributions were more concentrated than the empirical distributions. At the same time, the calibrated distributions were much less concentrated than the empirical distributions, reflecting the fact that participants were imperfect in estimating the shape and location of the distribution across categories. Both of these factors contributed to overprecision. Unsurprisingly, provision of information improved accuracy as measured by the concentration of the calibrated distribution; participants were less overprecise with the mode than without it. However, there is no evidence in our results that participants incorporated uncertainty about the distribution by having less concentrated subjective distributions. Even a strong

hint in the form of previously available information being taken away did not have a measurable impact on their level of confidence.

7. Experiment 3

In our third experiment, we sought to address remaining explanations for why people fail to properly account for epistemic uncertainty. To be well calibrated, judges must provide SPDs that are wider than they believe the empirical distribution to be. That requires them to appreciate that (a) the SPD is, in principle, different from one's best guess about the distribution, (b) the difference between them involves epistemic uncertainty, and (c) the appropriate response to epistemic uncertainty is to make SPDs wider. A failure on any of these three prerequisites could underlie judges' failure to respond to epistemic uncertainty.

In this experiment, we sought to make the distinction between judgments of probability and judgments of concentration as clear as we could. Rather than elicit the entire belief distribution, we asked participants to focus on the single most likely bin for a randomly selected exemplar (e.g., the most likely bin for the commute time of a randomly selected working adult in Austin, Texas). We asked one group of participants (the probability condition) to choose the category that they believed was most likely for the exemplar, and then to estimate the chances that the exemplar would be in the chosen category. As in our previous studies, this estimate should take into account both the aleatory uncertainty (what proportion of the population is in the most likely category) and empirical uncertainty (the probability that some other bin is in fact more likely). We asked another group (the concentration condition) to estimate the percentage of exemplars in the most common category, *whichever category that happened to be*. Participants do not need to account for uncertainty about which category is empirically the most likely when answering this question. If, for example, the judge's best guess is that 40% of Austin commuters are in the most common category of commute times, she should answer 40% to this question, even if she is uncertain about which category is in fact most common.

Note that the difference between these conditions makes a strong normative prediction. Reports in the probability condition should be lower than in the concentration condition. In the case of zero epistemic uncertainty—that is, when the respondent is absolutely certain they know the modal category—the two might be the same. But so long as there is any epistemic uncertainty, the probability reported in the probability condition must be lower than the probability reported in the concentration condition. This prediction depends on participants understanding the experimental instructions. As we explain in the next subsection, we took a number of additional steps to ensure that people understood

the questions as we intended. We sought to reduce, in so far as we could, the possibility that participants merely misunderstood what we were asking for.

If participants do distinguish between probability and concentration questions, do they also recognize that epistemic uncertainty is relevant to the difference? We gave some participants an *epistemic prompt*. Specifically, we asked them how confident they were that they correctly chose the most common category (probability condition) or *could* correctly choose the most common category if asked to do so (concentration condition). If such a direct prompt has no effect on judgments of either probability or concentration, it would suggest that judges do not see epistemic uncertainty as relevant to either type of question. If the prompt does affect either type of judgment, then we can observe whether the difference resembles the normative difference. That will provide evidence on whether people have any valid intuition about how to respond to combinations of epistemic and aleatory uncertainty.

7.1 Method

7.1.1. Participants. Our pre-registered research plan (https://osf.io/7b6m5/) called for a sample size of 960 participants from the ROI Rocket – ClearVoice online research panel. For this study, we wanted to ensure, to the extent possible, that participants would understand the distinction between probability and concentration and be comfortable working with numerical information. All potential participants were advised in advance that they would have to pass a math quiz to continue to the main study. Only those who passed were randomly assigned to a condition; those who failed were shown their responses along with the correct answers and informed that they could not continue. About 60% of potential participants (1,682 out of 2,720) passed the test and were randomly assigned to a condition. An additional 131 participants in the concentration condition, and 140 in the probability condition, dropped out during the course of the study. Our pre-registered research plan set quotas for each cell (a unique stimulus-condition combination), so prior to analysis we randomly dropped participants in groups where quotas were surpassed until the cell size was as required. In addition to a standard base payment of $0.50 from the survey company, participants also received an average bonus of $0.87 for accuracy. Demographic information was available for 97% of the participants. Of these, 65% were female; their average age was 49.3 (*S.D.* = 11.8).

7.1.2. Design. Each participant made probability estimate in response to six questions from the same set of domains and cities as in Experiment 2. Participants were evenly divided across the four between-subjects conditions in a 2 (estimate type: probability vs. concentration) x 2 (prompt: present vs. absent) design. Among those who received a prompt, half reported how confident they were that

that they had correctly chosen the most common category *before* they provided each estimate and half were asked *after*. Analyses that included the timing of the prompt showed no effects of that variable, so the analyses we report here collapse across the two timings.

We created sets of questions using a Latin Square for the order of domains, and six different cities were randomly chosen for each order. We repeated this process 20 times using the same original square of domain orders but pairing with new groups of cities. This resulted in 120 unique sets of six questions; later, each participant would receive one of these sets. The total collection of 720 questions (120 sets times 6 questions per set) featured each of the 40 candidate cities in 18 of these sets, and no set included the same city more than once.

7.1.3. Materials and procedures. The materials included screening questions, training questions, and estimate questions. The screening questions comprised five multiple-choice questions to assess numeracy (e.g., "Take 20% of 100, and then 50% of that. What do you get?"). Potential participants were required to answer at least four of five questions correctly. Those who did so received instructions appropriate to their condition, along with three questions to make sure that they understood the basic concepts of either probability or concentration and understood all the elements of the procedure. After answering each of these training questions, participants saw the correct answers alongside their own. (Consistent with our pre-registered research plan, participants were included in the sample regardless of their answers to these three questions.) Participants read that they should try to be as accurate as possible, and that they could earn up to an additional $0.20 cents for each of the three questions based on their accuracy.[4] The screening and training questions are available in the supplemental materials (https://osf.io/dt7cq/).

The estimate questions were like those used in Study 2, except that we reorganized the ranges of values in five of the six domains so that each question would have five total categories with three middle categories of equal width. For example, the ranges for commute times were < 15, 15-29, 30-44, 45-59, and ≥ 60 minutes. These modifications served to simplify the task for participants and to make results for different questions more comparable. Since the education domain is categorical, we simply

[4] Analogous to the payment scheme in Experiment 1, bonus payments were calculated for each question using an incentive-compatible extension of the Brier score. For each item, participants provided a probability judgment $\hat{p}$ (for either *the category they chose*, or *the most likely category, whichever category that happens to be*, in the probability and concentration conditions, respectively). We used the true probability p for that particular question to calculate an expected Brier score according to $EB = p(1-\hat{p})^2 + (1-p)\hat{p}^2$. The bonus earned was $20(1 - EB)$ cents for each of the six items.

reduced the number of categories to five (e.g., two of the original categories were combined into "Did not complete high school").

In the concentration condition, the estimate questions took this form: "In Atlanta, Georgia, what is the percentage of adult workers in the *most common* category of commuting times, whichever category that happens to be?" (Emphases were included in the actual stimuli.) Participants responded by selecting one of 21 radio buttons labeled 0% to 100% in increments of 5%. Those who received a prompt for this question were asked this either before or after each concentration question: "How confident are you that you could correctly choose the most common category of commuting times in Atlanta, Georgia." They responded by selecting one of five confidence levels, ranging from "Not at all confident" to "Extremely confident."

In the probability condition, estimates were elicited using a two-part question, such as, "In your judgment, which of the five commuting time categories is the *most common* one in Atlanta, Georgia?," followed by, "We're going to select a working adult in Atlanta, Georgia at random. What are the chances that this person was in the commuting time category that you chose?" They responded using the same 0%-100% scale as in the concentration condition. Participants who received a prompt were also asked (either before or after the probability question), "How confident are you that you correctly chose the *most common* commuting time category in Atlanta, Georgia?" They responded on the same 5-point confidence scale as described previously.

7.2 Results

We begin with an analysis of *estimates*, either of probability or of concentration, depending on the condition. We analyzed the data with a mixed model, with estimate type (probability vs. concentration) and prompt (present vs. absent) as the between-subject factors, and domain as within-subject factors. Results are shown in Table 4 and analyses in Table 5. If participants complied with Bayesian norms, estimates should be lower for probability than for concentration. We observe a small main effect of estimate type in that direction, but it is not statistically significant. If judges are more Bayesian when prompted to think of epistemic uncertainty, that will lead them to lower their estimates of probability, but not concentration. Thus, we would see a main effect of Prompt and an Estimate Type x Prompt interaction. We find no hint of either effect. Naturally, estimates differ from one domain to another. However, we did not predict the Type x Prompt interaction. For four of the five domains, assessments of concentration are slightly higher than assessments of probability; the reverse is true for the temperatures domain.

Table 4. Assessed probability mass vs. actual, collapsed across prompt conditions

	Concentration Condition: Mass of Maximum Category			Probability Condition Mass of Chosen Category		
Topic	Reported	Actual	Sub/Super-concentration	Reported	Actual	Over-precision
Ages	0.375	0.326	0.049	0.364	0.275	0.089
Commutes	0.430	0.434	-0.004	0.412	0.257	0.155
Education	0.426	0.297	0.128	0.409	0.247	0.161
Home Prices	0.438	0.425	0.013	0.408	0.278	0.129
Incomes	0.429	0.506	-0.077	0.404	0.281	0.124
Temperatures	0.446	0.471	-0.025	0.461	0.344	0.117
Mean	**0.424**	**0.410**	**0.014**	**0.410**	**0.280**	**0.129**

Table 5. Repeated measures ANOVA results for Experiment 3

Effect	effect df	adjusted effect df	error df	adjusted error df	F	Sig.	Partial Eta Squared
Mean	1		956		5555.37	<.001	.853
Domain	5	4.667	4780	4461.7	36.68	<.001	.037
Estimate Type	1				1.60	.201	.002
Prompt	1		592		0.00	.956	<.001
Type x Prompt	1				0.89	.346	.001
Domain x Type	5	4.667	4780	4461.7	3.17	.009	.003
Domain x Prompt	5	4.667	4780	4461.7	0.66	.644	.001
Domain x Type x Prompt	5	4.667	4780	4461.7	0.22	.946	<.001

Note. Adjusted degrees of freedom reflect the Greenhouse-Geisser correction for violations of sphericity.

Thus far, we have established that participants reported similar estimates on average regardless of whether they were evaluating concentration or probability. We next compare their estimates to the empirical distribution to determine their estimates' accuracy (concentration condition) and calibration (probability condition). In the concentration condition, we looked at the difference between estimated

and empirical concentration in a mixed 2 (prompt) x 6 (domain) ANOVA model. As in Experiments 1 and 2, there was a slight tendency toward super-concentration, M = .014, F(1, 478) = 3.19, p = .075, η_ρ^2 = 0.007. Not surprisingly, the domains differed from one another, F(4.31, 2058) = 91.04, p < .001, η_ρ^2 = 0.160.[5] However, as shown in Table 5, the tendency toward super-concentration was driven almost entirely by the education domain. The prompt had no bearing on super-concentration, F(1, 478) = 0.39, p = .531, η_ρ^2 = 0.001, nor was there a Prompt x Domain interaction, F(4.31, 2058) = 0.41, p = .802, η_ρ^2 = 0.001.

Next, using the data from the probability condition, we examined overprecision, the difference between the estimated and empirical probability mass in the category *chosen* as most common. Overall, participants overestimated the probability in the chosen category by 0.129, significantly different from zero, F(1, 478) =238.01, p = .001, η_ρ^2 = 0.332. Domains differed from one another in overprecision, F(4.56, 2180) = 10.14, p <.001, η_ρ^2 = 0.021. The prompt had no main effect, F(1, 478) = 0.73, p = .393, η_ρ^2 = 0.002, nor did it interact with domain, F(4.56, 2180) = .44, p = .807, η_ρ^2 = 0.001.

It appears from these findings that prompting people to consider their epistemic uncertainty had no effect. However, these are the net results across judges with different levels of confidence. When we take into account the judge's degree of confidence in correctly identifying the most common category, a different pattern emerges, shown in Figure 6. We performed a regression analysis on estimates with estimate type (concentration or probability), confidence, and domain as independent variables, along with the Estimate Type x Confidence interaction[6]. Estimates of concentration increased somewhat with greater confidence about which was the leading category, b = 0.024, *S.E.* = .012, t(479) = 2.00, p = .046. Estimates of probability increased with confidence more so, b = 0.097, with a significant Type x Confidence interaction, S.E. = .016, t(479) = 4.66, p < .001.

[5] Fractional degrees of freedom reflect the Greenhouse-Geisser correction for violations of sphericity.
[6] Type was dummy coded with 0 = concentration, domain was effect-coded, and clustered standard errors were used to deal with the possible interdependence of multiple responses from a single participant.

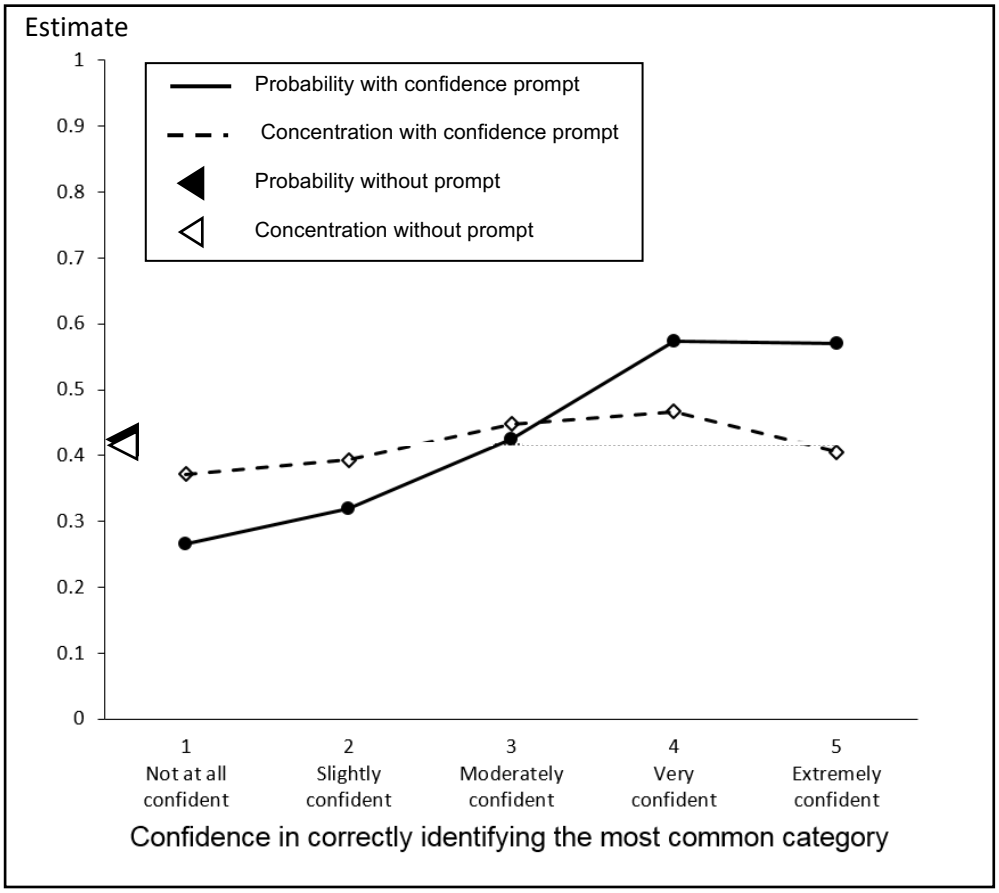

Figure 6. Relationships between confidence in identifying the most common category and estimates of (solid line) the probability of a randomly chosen item falling into the category believed most likely, or (dashed line) concentration, that is, the probability that a randomly chosen item falling in the most likely category, whichever category that is. Confidence in identifying the most likely category is not assessed in the no-prompt conditions (triangles).

The results make clear that participants' intuitions comply with normative prescriptions in two key respects: Their subjective probabilities are lower with greater epistemic uncertainty (i.e., with lower confidence about which is the most popular category, r = .336) and their estimates of the concentration are not sensitive to that variable. However, their intuitions are wrong in a way that explains the lack of overall difference between the two types of judgment. Judges violate the principle that the effect of epistemic uncertainty is unidirectional—probability judgments should only be lower than concentration judgments, or at most equal in the case of zero epistemic uncertainty. That is, the solid line should always be below the dashed line in Figure 6. Instead, a spotlight analysis shows that, although probability estimates were lower than concentration estimates when participants were "*Slightly confident*" ($b_{condition}$ = -.068, t(479) = -4.62, p < .001) and "*Not at all confident*" about which is the leading category ($b_{condition}$ = -.141, t(479) = -5.54, p < .001), the opposite was true when they were "*Very*

confident" ($b_{condition}$ = .078, *t(479)* = 2.72 *p* = .007) and "*Extremely confident*" ($b_{condition}$ = .151, *t*(479) = 4.49, *p* = .001). When, in the probability condition, participants estimate the chance of finding an instance in their chosen category, their probability estimates are well calibrated with empirical probabilities only for those who are "not at all confident."

7.3 Discussion

Experiment 3 provides further insight into where judges go wrong when epistemic and aleatory uncertainty combine. Subjective probability judgments should take into account both the aleatory distribution of possibilities in the population and one's uncertainty about what that distribution is. The presence of any degree of epistemic uncertainty means the judge's subjective probability of an event falling in the specific range believed to be most likely should be lower than the judge's best guess about the proportion of the population that is contained in its most likely range, *whichever that happens to be*. We find that judges do take epistemic uncertainty into account when estimating the probability of an event, at least when its presence is brought to their attention. With high epistemic uncertainty (i.e., low confidence about the population distribution), judges' subjective probabilities are indeed less concentrated than they believe the population to be. However, their intuitions about what to do about epistemic uncertainty seem to be wrong in one important respect. With high confidence (low epistemic uncertainty), judges' probabilities are *more* concentrated than their beliefs about the population. Intuitive judgments resemble an averaging of aleatory and epistemic uncertainty, rather than the unidirectional combination a statistician would prescribe.

8. General Discussion

Managers are regularly confronted with decision problems involving variables about which they have uncertainty. To make well-reasoned decisions, they must draw on their knowledge and beliefs about these variables to quantify the likelihood of different outcomes in the form of a subjective probability distribution. Often, this representation of lack of knowledge about the variable involves elements of both epistemic and aleatory uncertainty.

Most prior work on subjective confidence has focused only on epistemic uncertainty. Typically, there is a unique correct value (e.g., the year in which Mozart was born) and the only uncertainty arises from the judge's own lack of knowledge. However, people often face a second kind of uncertainty: aleatory, due to stochastic processes (e.g., how long it will take to get to work today). Many situations involve elements of both. Testing beliefs about aleatory uncertainty lay bare the distribution of outcomes and allow for tests of how precise or concentrated reported belief distributions are, relative to the true distribution of possible outcomes.

8.1 Empirical contributions

Our work provides insight into how judges take account of epistemic and aleatory uncertainty in thinking about a distribution of possible occurrences. Judges have imperfect knowledge about the distribution of values within any class of items or events, be it commute times or temperatures or prices. Their best guesses about these distributions are never exactly the right shape and size, and in exactly the right location. And, of course, judges do not know exactly what their errors may be. These sources of epistemic uncertainty mean that they should not try to match their subjective probability distributions to the concentration of the empirical distributions. Rather, well-calibrated judges should spread out their SPDs to account for the range of possibilities for what the empirical distributions are.

Moore, Carter, and Yang (2015) report that subjective probability distributions are less concentrated than the underlying empirical distributions, yet still too concentrated to be well-calibrated because they are placed, or centered, so badly. Our results suggest that, in general, people may be even further off the mark. Across a variety of more familiar domains, we found that SPDs were, on average, slightly *more* concentrated than the corresponding empirical distributions. This effect of super-concentration is not consistent across domains, but the finding that judges' SPDs are overprecise is quite robust in our data.

We examine several different explanations for why this is the case. Across studies, we vary the level of epistemic uncertainty, sometimes very transparently, and we provide hints and cues to bring epistemic uncertainty to front of mind. None of those manipulations had any appreciable effect. It does not, however, appear to be the case that judges simply neglect to consider epistemic uncertainty. Yet if they are aware of it, why don't they use their epistemic uncertainty to make their judgments of subjective probability less concentrated? Bayesian principles are notoriously unintuitive, so perhaps judges don't know that epistemic uncertainty should affect estimates of distributions, or perhaps they don't know how. Our final experiment favors the latter explanation. We find that judges do apply their sense of epistemic uncertainty to their subjective probabilities, but they have incorrect intuitions about how to do so. They spread out their SPDs in the presence of high epistemic uncertainty, but they concentrate them when epistemic uncertainty is low. Roughly speaking, judges seem to average different sources of uncertainty (aleatory and epistemic) rather than aggregating them (Soll 1999).

This interpretation comports with Tannenbaum et al. (2017), who show that people map epistemic and aleatory uncertainty into the 0-1 probability scale differently. Pure epistemic uncertainty gets mapped into the full scale—people tend to use extreme judgments when they perceive that they have complete knowledge about a situation. In contrast, pure aleatory uncertainty gets mapped into

more moderate probabilities that reflect perceptions of randomness. Tannenbaum et al. suggest that people employ an intermediate mapping (i.e., an average) for situations that involve a mixture of epistemic and aleatory uncertainty. We believe that this mapping error, combined with a tendency to overestimate one's knowledge about the distribution, is a major contributor toward overprecision.

8.2 Methodological contributions

We introduce a new method for characterizing the concentration of a probability distribution which allows a researcher to evaluate both (a) the concentration of a subjective distribution relative to the empirical distribution and (b) whether the subjective distribution is overprecise—too concentrated to achieve good calibration. These assessments rely on comparisons between three Gini coefficients, each of which measures the extent to which the probabilities from a particular distribution coalesce around a specific set of outcomes rather than being spread evenly across all outcomes. Comparisons of subjective, empirical, and calibrated Gini coefficients provide a novel viewpoint on the relationship between the concentration and calibration of a subjective distribution, offering insights into the causes of overprecision that are not available from existing metrics such as absolute deviations and interval hit rates. Furthermore, each Gini coefficient is measured in units of probability, thereby allowing for standardized comparisons across different distributions with different units, which facilitates analysis of judgments from a variety of domains.

8.1 Limitations and future directions

Our online samples provided greater diversity and larger sample sizes than would be possible in a laboratory. However, as with any online sample, there is room to be concerned about both their numerical sophistication and their motivation to be accurate. We attempted to address concerns about participants' numerical sophistication by screening participants with tests of numeracy. However, we cannot be certain that the screening ruled out all relevant misunderstandings of necessary numerical concepts. We attempted to motivate accuracy by providing monetary incentives that rewarded accurate responses. Although the size of these incentives was in line with recent norms for online participants, we cannot be certain that those modest monetary incentives were sufficient to insure strong accuracy motivation. It would be interesting to examine the possible effects of larger incentives and more training in mathematics and statistics. That said, we are interested in how probability judgments are made in the general population, with an ordinary range of skills and effort levels.

Like much of the prior literature, our experiments utilize assessments of probability distributions (Experiments 1 and 2) and numerical probabilities (Experiment 3). It has long been established that ordinary people misunderstand numerical probabilities (e.g., Fischhoff, 1991), and it is unlikely that

many people naturally think of uncertainty in terms of probability distributions. Thus, the elicitation methods we, and many previous investigators, use are unfamiliar to participants. Given how many decisions in life, from investment to clothing choices, depend on understanding probability and hedging risks, important questions remain about whether behavioral measures of certainty, such as choices under risk, might reveal more accurate intuitions (Mamassian, 2008; Mannes & Moore, 2013), and whether domain experts familiar with assessing risk are susceptible to the same errors when combining aleatory and epistemic uncertainty.

We have focused in this paper on the difficulty people have in combining aleatory and epistemic uncertainty. The distinction may also shed light on other findings in the overprecision literature, which future research should investigate. For example, there is less overprecision when participants learn by experience rather a description of events (Camilleri and Newell 2019), and when participants forecast future values (e.g., stock prices) based on time series as opposed to answering general knowledge questions (Budescu and Du 2007). In both cases, the direct observation of data is likely to favor an aleatory representation of the problem, which based on our findings should reduce but not eliminate overprecision. Future research might also investigate the implications for the accuracy-informativeness tradeoff (Yaniv and Foster 1995), which says that people avoid very wide confidence intervals because they are unhelpful to the listener (e.g., "90% confident that travel time is between 3 hours and 5 days"). Epistemic representations seem more flexible in assessing how much one knows, and are therefore more likely to favor informativeness over accuracy.

8.3 Conclusions

Our work touches on fundamental questions about how people know what they know. We find that people's estimates about a range of possible outcomes are systematically different from what norms suggest. Their probability judgments are overprecise, meaning that people underestimate the magnitude of their errors (Soll and Klayman, 2004). Subjective probability distributions do not properly reflect the degree of (in)accuracy in the judge's knowledge. We can easily imagine our participants objecting to our characterization of them. How could we expect them to know what they do not know? And yet, an appropriate level of confidence requires the application of exactly that kind of metacognition. How to do so is not at all obvious. Statistical reality demands that uncertainty in the placement of a distribution widen the distribution of possible outcomes, but this reality is not intuitively obvious to most people, at least not to our participants. And people are unlikely to get the kind of explicit, timely, and plentiful feedback needed to learn that fact through first-hand experience. Calibra, the perfectly calibrated judge, is imaginary precisely because the conditions permitting perfect

calibration are imaginary. As long as there are things people do not know, it will be difficult for them to fully take the lack of that unknown information into account when calibrating their confidence judgments.

Overconfidence in subjective probability distributions poses a key challenge for managerial decision making. Our analysis of variables involving both epistemic and aleatory uncertainty reveals novel insights about why and how judges' subjective probability distributions are overprecise. Future research should build on these findings to improve judgments about uncertain quantities. Identifying ways to reduce overconfidence remains a fundamental question, and we hope that our work inspires new approaches that are effective at overcoming this pervasive bias.

9. References

Alpert M, Raiffa H (1982) A progress report on the training of probability assessors. Kahneman D, Slovic P, Tversky A, eds. *Judgm. Uncertain. Heuristics Biases*. (Cambridge University Press, Cambridge).

Bernasco W, Steenbeek W (2017) More places than crimes: Implications for evaluating the law of crime concentration at place. *J. Quant. Criminol.* 33(3):451–467.

Brier GW (1950) Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* 78(1):1–3.

Budescu DV, Du N (2007) The coherence and consistency of investors' probability judgments. *Manag. Sci.* 53(11):1731–1745.

Camilleri AR, Newell BR (2019) Better calibration when predicting from experience (rather than description). *Organ. Behav. Hum. Decis. Process.* 150:62–82.

Collins NR, Preston LE (1961) The size structure of the largest industrial firms, 1909-1958. *Am. Econ. Rev.* 51(5):986–1011.

Deltas G (2003) The small-sample bias of the Gini coefficient: results and implications for empirical research. *Rev. Econ. Stat.* 85(1):226–234.

Du N, Budescu DV (2018) How (Over) Confident Are Financial Analysts? *J. Behav. Finance* 19(3):308–318.

Fischhoff B (1991) Value elicitation: Is there anything in there? *Am. Psychol.* 46(8):835–847.

Fox CR, Ülkümen G (2011) Distinguishing two dimensions of uncertainty. Brun W, Keren G, Kirkebøen G, Montgomery H, eds. *Perspect. Think. Judg. Decis. Mak.* (Universitetsforlaget, Oslo), 21–35.

Gastwirth JL (1972) The estimation of the Lorenz curve and Gini index. *Rev. Econ. Stat.* 54(3):306–316.

Gigerenzer G (1994) Why the distinction between single-event probabilities and frequencies is important for psychology (and vice versa). Wright G, Ayton P, eds. *Subj. Probab.* (Wiley, Oxford), 129–161.

Gini C (1912) Variabilità e mutabilità. *Repr. Mem. Metodol. Stat. Ed Pizetti E Salvemini T Rome Libr. Eredi Virgilio Veschi*.

Glaser M, Weber M (2007) Overconfidence and trading volume. *Geneva Risk Insur. Rev.* 32:1–36.

Goldstein DG & Rothschild D (2014) Lay understanding of probability distributions. *Judgm. Decis. Mak.*, 9(1).

Haran U, Moore DA, Morewedge CK (2010) A simple remedy for overprecision in judgment. *Judgm. Decis. Mak.* 5(7):467–476.

Howell WC, Burnett SA (1978) Uncertainty measurement: A cognitive taxonomy. *Organ. Behav. Hum. Perform.* 22(1):45–68.

Jain K, Mukherjee K, Bearden JN, Gaba A (2013) Unpacking the Future: A Nudge Toward Wider Subjective Confidence Intervals. *Manag. Sci.*

Kahneman D, Tversky A (1982) On the study of statistical intuitions. *Cognition* 11(2):123–141.

Klayman J, Soll JB, Gonzalez-Vallejo C, Barlas S (1999) Overconfidence: It depends on how, what, and whom you ask. *Organ. Behav. Hum. Decis. Process.* 79(3):216–247.

Lichtenstein S, Fischhoff B, Phillips LD (1982) Calibration of probabilities: The state of the art in 1980. Kahneman D, Slovic P, Tversky A, eds. *Judgm. Uncertain. Heuristics Biases*. (Cambridge University Press, Cambridge, England), 306–333.

Lorenz MO (1905) Methods of measuring the concentration of wealth. *Publ. Am. Stat. Assoc.* 9(70):209–219.

Mamassian P (2008) Overconfidence in an objective anticipatory motor task. *Psychol. Sci.* 19(6):601–606.

Mannes AE, Moore DA (2013) A behavioral demonstration of overconfidence in judgment. *Psychol. Sci.* 24(7):1190–1197.

Moore DA, Carter A, Yang HHJ (2015) Wide of the mark: Evidence on the underlying causes of overprecision in judgment. *Organ. Behav. Hum. Decis. Process.* 131:110–120.

Moore DA, Healy PJ (2008) The trouble with overconfidence. *Psychol. Rev.* 115(2):502–517.

Nisbett RE, Krantz DH, Jepson C, Kunda Z (1983) The use of statistical heuristics in everyday inductive reasoning. *Psychol. Rev.* 90(4):339–363.

Nisbett RE, Kunda Z (1985) Perception of social distributions. *J. Pers. Soc. Psychol.* 48(2):297.

Peterson DK, Pitz GF (1988) Confidence, uncertainty, and the use of information. *J. Exp. Psychol. Learn. Mem. Cogn.* 14(1):85.

Soll JB (1999) Intuitive theories of information: Beliefs about the value of redundancy. *Cognit. Psychol.* 38:317–346.

Soll JB, Klayman J (2004) Overconfidence in interval estimates. *J. Exp. Psychol. Learn. Mem. Cogn.* 30(2):299–314.

Tannenbaum D, Fox CR, Ülkümen G (2017) Judgment extremity and accuracy under epistemic versus aleatory uncertainty. *Manag. Sci.* 63(2):497–518.

Teigen KH (1994) Variants of subjective probabilities: Concepts, norms, and biases. Wright G, Ayton P, eds. *Subj. Probab.* (Wiley, Oxford), 211–238.

Teigen KH, Jørgensen M (2005) When 90% confidence intervals are 50% certain: On the credibility of credible intervals. *Appl. Cogn. Psychol.* 19(4):455–475.

Ülkümen G, Fox CR, Malle BF (2016) Two dimensions of subjective uncertainty: Clues from natural language. *J. Exp. Psychol. Gen.* 145(10):1280–1297.

Yaniv I, Foster DP (1995) Graininess of judgment under uncertainty: An accuracy-informativeness trade-off. *J. Exp. Psychol. Gen.* 124(4):424–32.